

THE LASSO FOR GENERIC DESIGN MATRICES AS A FUNCTION OF THE RELAXATION PARAMETER

STÉPHANE CHRÉTIEN AND SÉBASTIEN DARSE

ABSTRACT. The LASSO is a variable subset selection procedure in statistical linear regression based on ℓ_1 penalization of the least-squares operator. Its behavior crucially depends, both in practice and in theory, on the ratio between the fidelity term and the penalty term. We provide a detailed analysis of the fidelity vs. penalty ratio as a function of the relaxation parameter. Our study is based on a general position condition on the design matrix which holds with probability one for most experimental models. Along the way, the proofs of some well known basic properties of the LASSO are provided from this new generic point of view.

1. INTRODUCTION

1.1. Problem statement and main results. The well-known standard Gaussian linear model in statistics reads $y = X\beta + z$, where X denotes a $n \times p$ design matrix, $\beta \in \mathbb{R}^p$ is an unknown parameter and the components of the error z are assumed i.i.d. with normal distribution $\mathcal{N}(0, \sigma^2)$. Let us briefly recall some basic notations. For $I \subset \{1, \dots, p\}$, $|I|$ denotes the cardinal of I . For $x \in \mathbb{R}^p$, we set $x_I = (x_i)_{i \in I} \in \mathbb{R}^{|I|}$. The usual scalar product is denoted by $\langle \cdot, \cdot \rangle$. The notations for the norms on vectors and matrices are also standard: for any vector $x = (x_i) \in \mathbb{R}^N$,

$$\|x\|_2^2 = \sum_{1 \leq i \leq N} x_i^2; \quad \|x\|_1 = \sum_{1 \leq i \leq N} |x_i|; \quad \|x\|_\infty = \sup_{1 \leq i \leq N} |x_i|.$$

For any matrix A , we denote by A^t its transpose. For $I \subset \{1, \dots, p\}$, and a matrix X , we denote by X_I the submatrix whose columns are indexed by I .

The case where p is much larger than n has been the subject of an intense recent study. This problem is of course not solvable for any β but it has been discovered that if β is sufficiently sparse, then the solution of

$$(1.1) \quad \hat{\beta}_\lambda \in \operatorname{argmin}_{b \in \mathbb{R}^p} \frac{1}{2} \|y - Xb\|_2^2 + \lambda \|b\|_1,$$

called the LASSO estimator of β , is sometimes also sparse and close to β . The acronym LASSO, due to [11], stands for Least Absolute Shrinkage and Selection Operator, and stems from the fact that the ℓ_1 -norm penalty shrinks the components of the standard least-squares estimator $\hat{\beta}$. Some components are shrunk to the point of setting them to zero, hence implying automatic selection of the remaining nonzero components as good predictors for the experiments under study. We refer the interested reader to [5] and [9] for an overview of the relationships between sparsity and statistics, and sparsity promoting penalizations of the least-squares criterion. Recent results

concerning the LASSO and extensions to other statistical models and penalizations strategies may be found in [3], [1], [6] and [12] for instance.

Under the assumption that the columns of X are sufficiently "uncorrelated", several authors were able to prove that, with high probability, the ℓ_2 -norm of $X(\hat{\beta} - \beta)$ is of the same order of magnitude as the ℓ_2 -norm of $X(\tilde{\beta} - \beta)$ for an oracle $\tilde{\beta}$. It may even perform as well as an "oracle". For instance, the oracle proposed in [6] is a solution of

$$\tilde{\beta}_\lambda \in \underset{b \in \mathbb{R}^p, b_T = 0}{\operatorname{argmin}} \frac{1}{2} \|y - X_T b_T\|_2^2 + \lambda \operatorname{sgn}(\beta_T)^t b_T,$$

where T is the index set of the non-zero components of β . The term "oracle" is often used to emphasize that the support of β is usually unknown ahead of time. Under stronger assumptions it was further proven in [4] and [6] that the support and sign pattern of β can be recovered exactly with high probability. A very efficient algorithm, based on Nesterov's method, for solving the LASSO estimation problem is described in [2].

A central quantity in the numerical analysis of the LASSO is the ratio

$$\Gamma = \frac{\lambda \|\hat{\beta}_\lambda\|_1}{\|y - X\hat{\beta}_\lambda\|_2^2}.$$

As is well known to both practitioners and theoreticians, severe problems occur when Γ is either very small or very large. The present paper provides a detailed analysis of Γ as a function of λ . The proofs rely on a general position condition which holds with probability one for most random design matrix models. Along the way, we prove from this generic view point several results on the LASSO estimator which seem to belong to the folklore: uniqueness, continuity and piecewise affine parametrization as a function of λ . Our main result states that there exists $\tau > 0$ such that Γ is decreasing on $(0, \tau]$ with $\Gamma(\tau) = 0$, and that $\|y - X\hat{\beta}_\lambda\|_2$ is increasing on $(0, \tau]$.

1.2. The General Position Condition. Our main assumption on the design matrix X is the following.

Assumption 1.1. (General Position Condition for X) *For all supports $S \neq S' \subset \{1, \dots, n\}$ and all $(\varepsilon_S, \varepsilon_{S'}) \in \{-1, 1\}^{|S|} \times \{-1, 1\}^{|S'|}$ such that X_S and $X_{S'}$ are non-singular, we have*

$$(1.2) \quad \varepsilon_S (X_S^t X_S)^{-1} \varepsilon_S \neq \varepsilon_{S'}^t (X_{S'}^t X_{S'})^{-1} \varepsilon_{S'}$$

$$(1.3) \quad \varepsilon_S (X_S^t X_S)^{-1} (X_S^t X_{S'}) (X_{S'}^t X_{S'})^{-1} \varepsilon_{S'} \neq |\varepsilon_S (X_S^t X_S)^{-1} \varepsilon_S|.$$

Since $S \neq S'$, this property clearly holds with probability one if the entries of X are independent and have an absolutely continuous density with respect to the Lebesgue measure. This is a generic situation in statistics where the covariate measurements are usually corrupted by some noise. In the case of a more general type of design, we believe that this definition could easily be generalized so as to guarantee that (1.2) fails with probability at most of the order $p^{-\alpha}$ or is automatically satisfied for a carefully chosen deterministic design. A similar property, called Unicity Condition (UC) was proposed in [7] for the problem of finding the sparsest solution of a linear system with application to the field of compressed sensing.

1.3. Plan of the paper. Section 2 recalls the optimality conditions associated to the LASSO. In Section 3, we study the standard LASSO estimator of β as a function of λ . In particular, various continuity and monotonicity properties of some important functions of $\hat{\beta}_\lambda$ using the General Position Condition assumption only are established. Based on these results, we prove in Section 4 our main result Theorem 4.2.

1.4. Additional notations. The set of symmetric real matrices is denoted by \mathbb{S}_n . For any matrix A in $\mathbb{R}^{d_1 \times d_2}$, we denote by $\|A\|$ the operator norm of A . The maximum (resp. minimum) singular value of A is denoted by σ_{\max} (resp. $\sigma_{\min}(A)$). Recall that $\sigma_{\max}(A) = \|A\|$ and $\sigma_{\min}(A)^{-1} = \|A^{-1}\|$. We use the Loewner ordering on symmetric real matrices: if $A \in \mathbb{S}_n$, $0 \preceq A$ is equivalent to saying that A is positive semi-definite, and $A \preceq B$ stands for $0 \preceq B - A$.

For any vector $b \in \mathbb{R}^p$, b^+ (resp. b^-) denotes its non-negative (resp. non-positive) part, i.e. $b = b^+ - b^-$, with $b_j^+, b_j^- \geq 0$.

For a given support $S \subset \{1, \dots, n\}$, we denote the range of X_S by V_S and the orthogonal projection onto V_S by \mathbf{P}_{V_S} . Recall that

$$\mathbf{P}_{V_S} = X_S(X_S^t X_S)^{-1} X_S^t.$$

The support of $\hat{\beta}_\lambda$ is denoted by \hat{T}_λ . For the sake of notational simplicity, we write

$$(1.4) \quad \hat{\beta}_{\hat{T}_\lambda} := \left(\hat{\beta}_\lambda \right)_{\hat{T}_\lambda}.$$

2. OPTIMALITY CONDITIONS

In this section, we review the standard optimality conditions for the LASSO estimator. A necessary and sufficient optimality condition in (1.1) is that

$$(2.5) \quad 0 \in \partial \left(\frac{1}{2} \|y - X \hat{\beta}_\lambda\|_2^2 + \lambda \|\hat{\beta}_\lambda\|_1 \right),$$

where ∂ denotes the sub-differential, which is equivalent to the existence of g_λ in $\partial \|\cdot\|_1$ at $\hat{\beta}_\lambda$ such that

$$(2.6) \quad -X^t(y - X \hat{\beta}_\lambda) + \lambda g_\lambda = 0.$$

On the other hand, the sub-differential of $\|\cdot\|_1$ at $\hat{\beta}_\lambda$ is defined by

$$\partial \|\cdot\|_1(\hat{\beta}_\lambda) = \left\{ \gamma \in \mathbb{R}^p, \gamma_{\hat{T}_\lambda} = \text{sgn}(\hat{\beta}_{\hat{T}_\lambda}) \text{ and } \|\gamma_{\hat{T}_\lambda^c}\|_\infty < 1 \right\}.$$

Thus, using the fact that $y = X\beta + z$, we may easily conclude that a necessary and sufficient condition for optimality in (1.1) is the existence of a vector g_λ , satisfying $g_{\hat{T}_\lambda} = \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})$ and $\|g_{\hat{T}_\lambda^c}\|_\infty < 1$, and such that

$$(2.7) \quad X^t(y - X \hat{\beta}_\lambda) = \lambda g_\lambda.$$

The following corollary is a direct but important consequence of these previous preliminary remarks.

Corollary 2.1. *A necessary and sufficient condition for a given random vector \mathbf{b} with support \mathbf{T} to simultaneously satisfy the two following conditions:*

- (1) $\mathbf{b} = \hat{\beta}_\lambda$,

(2) \mathbf{b} has the same support T and sign pattern $\text{sgn}(\beta_T)$ as β is that

$$(2.8) \quad X_T^t(y - X\mathbf{b}) = \lambda \text{sgn}(\beta_T)$$

$$(2.9) \quad \|X_{T^c}^t(y - X\mathbf{b})\|_\infty < \lambda.$$

Proof. The fact that (2.8) and (2.9) are necessary is a straightforward consequence of (2.7). Conversely, assume that (2.8) and (2.9) hold. Set

$$(2.10) \quad \mathbf{g} = \frac{1}{\lambda} X^t(y - X\mathbf{b}).$$

Using (2.6), we deduce that \mathbf{g} belongs to $\partial\|\cdot\|_1(\mathbf{b})$ and that the support of \mathbf{b} is exactly the set $\mathbb{T} = \{j \in \{1, \dots, p\}, |\mathbf{g}_j| = 1\}$. On the other hand, we have that

$$(2.11) \quad \mathbf{g} = \text{sgn}(\beta_T)$$

$$(2.12) \quad \|\mathbf{g}\|_\infty < 1,$$

and we may deduce that \mathbf{g} is at the same time in the sub-differential of any vector b in \mathbb{R}^p with same support and sign pattern as β . Therefore, we have

$$(2.13) \quad T = \{j \in \{1, \dots, p\}, |\mathbf{g}_j| = 1\} = \mathbb{T},$$

and we conclude that β and \mathbf{b} have the same support. Moreover, the index set T^+ of the positive components of β and the index set \mathbb{T}^+ of the positive components of \mathbf{b} satisfy

$$(2.14) \quad T^+ = \{j \in \{1, \dots, p\}, \mathbf{g}_j = 1\} = \mathbb{T}^+.$$

The same argument implies that the index set T^- of the negative components of β equals the index set \mathbb{T}^- of the negative components of \mathbf{b} . To sum up, β and \mathbf{b} have the same support and sign pattern and the proof is completed. This moreover implies that (2.8) and (2.9) are the optimality conditions for (1.1) and we obtain that $\mathbf{b} = \hat{\beta}$ as announced. \square

3. THE LASSO ESTIMATOR AS A FUNCTION OF λ

This section establishes various continuity and monotonicity properties of some important functions of $\hat{\beta}_\lambda$ using the General Position Condition assumption only.

The following notations will be useful. Define \mathcal{L} as the cost function:

$$(3.15) \quad \mathcal{L} : \begin{cases} \mathbb{R}_+^* \times \mathbb{R}^p & \longrightarrow & \mathbb{R}_+ \\ (\lambda, b) & \longmapsto & \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1, \end{cases}$$

and for all $\lambda > 0$,

$$(3.16) \quad \theta(\lambda) = \inf_{b \in \mathbb{R}^p} \mathcal{L}(\lambda, b).$$

3.1. More on the estimator $\hat{\beta}_\lambda$. We begin with the following useful characterization of the LASSO estimators. For any $w \in \mathbb{R}^p$, let us introduce

$$(3.17) \quad \mathcal{P}(w) = \underset{b \in \mathbb{R}^p, Xb=Xw}{\operatorname{argmin}} \|b\|_1.$$

Lemma 3.1. *A vector $\hat{\beta}_\lambda$ is a solution of (1.1) if and only if $\hat{\beta}_\lambda \in \mathcal{P}(\hat{\beta}_\lambda)$.*

Proof. Let $\hat{\beta}_\lambda$ be a solution of (1.1). Let $\tilde{\beta}_\lambda \in \mathcal{P}(\hat{\beta}_\lambda)$. Then, we have

$$(3.18) \quad \|\tilde{\beta}_\lambda\|_1 \leq \|\hat{\beta}_\lambda\|_1.$$

On the other hand, the definition of $\hat{\beta}_\lambda$ implies that

$$(3.19) \quad \frac{1}{2}\|y - X\tilde{\beta}_\lambda\|_2^2 + \lambda\|\tilde{\beta}_\lambda\|_1 \geq \frac{1}{2}\|y - X\hat{\beta}_\lambda\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_1.$$

Moreover, since $X\tilde{\beta}_\lambda = X\hat{\beta}_\lambda$, we have that

$$(3.20) \quad \frac{1}{2}\|y - X\tilde{\beta}_\lambda\|_2^2 = \frac{1}{2}\|y - X\hat{\beta}_\lambda\|_2^2,$$

and subtracting this equality to (3.19), we obtain that

$$\|\hat{\beta}_\lambda\|_1 \leq \|\tilde{\beta}_\lambda\|_1,$$

which, combined with (3.18), implies that

$$(3.21) \quad \|\hat{\beta}_\lambda\|_1 = \|\tilde{\beta}_\lambda\|_1.$$

This last equality together with (3.20) implies the desired result. \square

We now give a useful expression of $\hat{\beta}_\lambda$ in terms of λ and the submatrix of X indexed by \hat{T} .

Lemma 3.2. *For any $\lambda > 0$ such that $\hat{\beta}_\lambda \neq 0$, the matrix $X_{\hat{T}_\lambda}$ is non-singular and we have*

$$(3.22) \quad \hat{\beta}_{\hat{T}_\lambda} = (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} (X_{\hat{T}_\lambda}^t y - \lambda \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})).$$

Proof. Recall that the optimality conditions for the LASSO imply that

$$(3.23) \quad X_{\hat{T}_\lambda}^t (y - X_{\hat{T}_\lambda} \hat{\beta}_{\hat{T}_\lambda}) = \lambda \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda}).$$

Since $X_{\hat{T}_\lambda}$ is non-singular, we obtain (3.22). \square

The following Theorem establishes the existence and unicity of $\hat{\beta}_\lambda$ with support size less than or equal to n . In the sequel, $\hat{\beta}_\lambda$ will always refer to this solution.

Theorem 3.3. *Assume that Assumption 1.1 holds. Then, almost surely, for any $\lambda > 0$, the minimization problem (1.1) has a unique solution $\hat{\beta}_\lambda$ with support $\hat{T}_\lambda \subset \{1, \dots, p\}$ verifying*

$$(3.24) \quad |\hat{T}_\lambda| \leq n.$$

Proof. We first study the support of a possible solution $\hat{\beta}_\lambda$. Second, we derive (3.22), and eventually, we prove the uniqueness of $\hat{\beta}_\lambda$ under the general position condition.

Study of $\#\hat{T}$. Recall that b^+ (resp. b^-) be the non-negative (resp. non-positive) part of b , i.e. $b = b^+ - b^-$, with $b_j^+, b_j^- \geq 0$. Then, Lemma 3.1 above equivalently says that $\hat{\beta}_\lambda$ is a solution of (1.1) if and only if $\hat{\beta}_\lambda^+$ and $\hat{\beta}_\lambda^-$ are solutions of

$$(3.25) \quad \min_{b^+, b^- \in \mathbb{R}_+^p} \sum_{j=1}^p \{b_j^+ + b_j^-\} \text{ s.t. } Xb^+ - Xb^- = X\hat{\beta}_\lambda.$$

The remainder of the proof relies on linear programming theory and Assumption 1.1. Notice first that the solution set is compact due to the coercivity of the ℓ_1 -norm. Thus, the theory of linear programming [10] ensures that each extreme point of the solution set of (3.25) is completely determined by a "basis" B . In the present setting, for an extreme point $b^* = b^{*+} - b^{*-}$ of the solution set of (3.25), the associated basis B^* can be written (in a non-unique way) as $B^* = B^{*+} \cup B^{*-}$, $|B^*| = n$, and is such that

- (i) the square matrix $[X_{B^{*+}}, -X_{B^{*-}}]$ is non singular,
- (ii) $b_{B^{*c}}^* = 0$ and
- (iii) the couple $(b_{B^{*+}}^{*+}, b_{B^{*-}}^{*-})$ is uniquely determined by the system

$$(3.26) \quad X_{B^{*+}} b_{B^{*+}}^{*+} - X_{B^{*-}} b_{B^{*-}}^{*-} = X\hat{\beta}_\lambda,$$

(or equivalently, $X_{B^*} b_{B^*}^* = X\hat{\beta}_\lambda$).

An immediate consequence is that the support of b^* has cardinal at most n . Moreover, $b^* \in \mathcal{P}(b^*)$, and using Lemma 3.1, we deduce that b^* is a solution of (1.1). Therefore, we may assume without loss of generality that $\hat{\beta}_\lambda$ is an extreme point of $\mathcal{P}(\hat{\beta}_\lambda)$, with

$$\#\hat{T}_\lambda \leq n$$

and that $X_{\hat{T}_\lambda}$ is non-singular.

Uniqueness of $\hat{\beta}_\lambda$: first part. — We give two equations satisfied by λ and z in the case where uniqueness of the LASSO estimator fails.

Let $\hat{\beta}'_\lambda$ in \mathbb{R}^p be another solution of (1.1). Using the same reasoning as for $\hat{\beta}_\lambda$ in the end of the last paragraph, we may assume w.l.o.g. that the support \hat{T}'_λ of $\hat{\beta}'_\lambda$ has cardinal at most n and that $X_{\hat{T}'_\lambda}$ is non-singular. Convexity of the LASSO functional implies that the map

$$(3.27) \quad \phi : \begin{cases} [0, 1] & \longrightarrow \mathbb{R}_+ \\ t & \longmapsto \mathcal{L}(\lambda, (t\hat{\beta}_\lambda + (1-t)\hat{\beta}'_\lambda)) \end{cases}$$

is constant.

Notice that the term $\|\hat{\beta}'_\lambda + t(\hat{\beta}_\lambda - \hat{\beta}'_\lambda)\|_1$ is in fact piecewise affine on $(0, t)$. Set

$$\begin{aligned} \rho_\lambda &= \text{sgn}(\hat{\beta}_{\hat{T}_\lambda}) \\ \rho'_\lambda &= \text{sgn}(\hat{\beta}'_{\hat{T}'_\lambda}). \end{aligned}$$

Now, let $t^* > 0$ sufficiently small such that for all $t \in (0, t^*)$ the support of $\hat{\beta}'_\lambda + t(\hat{\beta}_\lambda - \hat{\beta}'_\lambda)$ is constant and equal to $\hat{T}_\lambda \cup \hat{T}'_\lambda$ and no sign change occurs.

Set

$$(3.28) \quad \rho = \operatorname{sgn} \left(\left(\widehat{\beta}'_{\lambda} + t (\widehat{\beta}_{\lambda} - \widehat{\beta}'_{\lambda}) \right)_{\widehat{T}_{\lambda} \cup \widehat{T}'_{\lambda}} \right).$$

Thus, for all $t \in (0, t^*)$,

$$\|\widehat{\beta}'_{\lambda} + t (\widehat{\beta}_{\lambda} - \widehat{\beta}'_{\lambda})\|_1 = \rho^t \widehat{\beta}'_{\lambda} + t \rho^t (\widehat{\beta}_{\lambda} - \widehat{\beta}'_{\lambda})$$

with

$$\rho_{\widehat{T}_{\lambda}} = \rho_{\lambda} \quad \text{and} \quad \rho_{\widehat{T}'_{\lambda}} = \rho'_{\lambda}$$

and we deduce that ϕ is a second order polynomial in the variable $t \in (0, t^*)$. Therefore, the coefficients corresponding to the quadratic and linear terms of ϕ must be zero. Developing the term $\frac{1}{2} \|y - X(t \widehat{\beta}_{\lambda} + (1-t) \widehat{\beta}'_{\lambda})\|_2^2$, we then obtain:

$$\begin{aligned} X_{\widehat{T}_{\lambda}} \widehat{\beta}_{\lambda} - X_{\widehat{T}'_{\lambda}} \widehat{\beta}'_{\lambda} &= 0 \\ y^t (X_{\widehat{T}_{\lambda}} \widehat{\beta}_{\lambda} - X_{\widehat{T}'_{\lambda}} \widehat{\beta}'_{\lambda}) + \lambda \rho^t (\widehat{\beta}_{\lambda} - \widehat{\beta}'_{\lambda}) &= 0, \end{aligned}$$

which is equivalent to

$$(3.29) \quad X_{\widehat{T}_{\lambda}} \widehat{\beta}_{\lambda} - X_{\widehat{T}'_{\lambda}} \widehat{\beta}'_{\lambda} = 0$$

$$(3.30) \quad \rho^t (\widehat{\beta}_{\lambda} - \widehat{\beta}'_{\lambda}) = 0.$$

Uniqueness of $\widehat{\beta}_{\lambda}$: second part. — As for $\widehat{\beta}_{\widehat{T}_{\lambda}}$, we write

$$(3.31) \quad \widehat{\beta}'_{\widehat{T}'_{\lambda}} = (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \left(X_{\widehat{T}'_{\lambda}}^t y - \lambda \operatorname{sgn}(\widehat{\beta}'_{\widehat{T}'_{\lambda}}) \right).$$

Replacing (3.22) and (3.31) into (3.29), we obtain

$$(3.32) \quad (\mathbf{P}_{\widehat{T}_{\lambda}} - \mathbf{P}_{\widehat{T}'_{\lambda}}) y - \lambda \left(X_{\widehat{T}_{\lambda}} (X_{\widehat{T}_{\lambda}}^t X_{\widehat{T}_{\lambda}})^{-1} \rho_{\lambda} - X_{\widehat{T}'_{\lambda}} (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \rho'_{\lambda} \right) = 0.$$

On the other hand, (3.30) gives

$$(3.33) \quad 0 = y^t \left(X_{\widehat{T}_{\lambda}} (X_{\widehat{T}_{\lambda}}^t X_{\widehat{T}_{\lambda}})^{-1} \rho_{\lambda} - X_{\widehat{T}'_{\lambda}} (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \rho'_{\lambda} \right) - \lambda \left(\rho_{\lambda}^t (X_{\widehat{T}_{\lambda}}^t X_{\widehat{T}_{\lambda}})^{-1} \rho_{\lambda} - (\rho'_{\lambda})^t (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \rho'_{\lambda} \right).$$

Setting

$$\begin{aligned} \eta_{\lambda} &= X_{\widehat{T}_{\lambda}} (X_{\widehat{T}_{\lambda}}^t X_{\widehat{T}_{\lambda}})^{-1} \rho_{\lambda} - X_{\widehat{T}'_{\lambda}} (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \rho'_{\lambda} \\ \zeta_{\lambda} &= \rho_{\lambda}^t (X_{\widehat{T}_{\lambda}}^t X_{\widehat{T}_{\lambda}})^{-1} \rho_{\lambda} - (\rho'_{\lambda})^t (X_{\widehat{T}'_{\lambda}}^t X_{\widehat{T}'_{\lambda}})^{-1} \rho'_{\lambda}, \end{aligned}$$

we obtain the system:

$$(3.34) \quad (\mathbf{P}_{\widehat{T}_{\lambda}} - \mathbf{P}_{\widehat{T}'_{\lambda}}) y - \lambda \eta_{\lambda} = 0$$

$$(3.35) \quad y^t \eta_{\lambda} - \lambda \zeta_{\lambda} = 0.$$

Notice that

$$\left(\mathbf{P}_{\widehat{T}_{\lambda}} - \mathbf{P}_{\widehat{T}'_{\lambda}}, \eta_{\lambda}, \zeta_{\lambda} \right) \in \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3,$$

where

$$\begin{aligned}\mathcal{F}_1 &= \{\mathbf{P}_S - \mathbf{P}_{S'}, S \neq S' \subset \{1, \dots, n\}\} \\ \mathcal{F}_2 &= \{X_S(X_S^t X_S)^{-1} \varepsilon_S - X_{S'}(X_{S'}^t X_{S'})^{-1} \varepsilon_{S'}, (S, S', \varepsilon_S, \varepsilon_{S'}) \in \mathcal{G}\} \\ \mathcal{F}_3 &= \{\varepsilon_S^t (X_S^t X_S)^{-1} \varepsilon_S - \varepsilon_{S'}^t (X_{S'}^t X_{S'})^{-1} \varepsilon_{S'} \mid (S, S', \varepsilon_S, \varepsilon_{S'}) \in \mathcal{G}\},\end{aligned}$$

with

$$\mathcal{G} = \{S \neq S' \subset \{1, \dots, n\}, (\varepsilon_S, \varepsilon_{S'}) \in \{-1, 1\}^{|S|} \times \{-1, 1\}^{|S'|}\}.$$

Therefore, (y, λ) is a solution of the finite set of equations

$$(3.36) \quad Q y - \lambda \eta = 0$$

$$(3.37) \quad y^t \eta - \lambda \zeta = 0,$$

when (Q, η, ζ) is running over $\mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$. This implies that

$$\left\{ \left(\hat{\beta}_\lambda, \lambda \right), \lambda > 0 \right\} \subset \bigcup_{j \in \mathcal{J}} E_j,$$

where \mathcal{J} is a finite set and the $E_j \subset \mathbb{R}^{n+1}$ are linear subspaces.

Let us now show that there is no E_j , $j \in \mathcal{J}$, containing a subspace of dimension n . Let us suppose that this is not the case, i.e. there exist two supports $S \neq S'$ and $(\eta, \zeta) \in \mathcal{F}_2 \times \mathcal{F}_3$ such that for all $y \in \mathbb{R}^n$,

$$(3.38) \quad (\mathbf{P}_S - \mathbf{P}_{S'})y = \frac{\eta \eta^t}{\zeta} y.$$

When the rank of $\mathbf{P}_S - \mathbf{P}_{S'}$ is different from 1, (3.38) cannot be satisfied for all $y \in \mathbb{R}^n$. Thus, we only have to focus on the case where the rank of $\mathbf{P}_S - \mathbf{P}_{S'}$ is 1, or equivalently, $|S \Delta S'| = 1$. We distinguish two cases. Either $W_S := V_S^\perp \cap V_{S'} \neq \{0\}$ or $W_S = \{0\}$:

- (i) If $W_S \neq \{0\}$, take $v \in W_S$, $v \neq 0$. Then $(\mathbf{P}_S - \mathbf{P}_{S'})v = -v$, and the only eigenvalue of $\mathbf{P}_S - \mathbf{P}_{S'}$ is -1 .
- (ii) If $W_S = \{0\}$, then $V_{S'} \subset V_S$ and so $W_{S'} := V_{S'}^\perp \cap V_S \neq \{0\}$. Hence, take a non-zero $v \in W_{S'}$. We now have $(\mathbf{P}_S - \mathbf{P}_{S'})v = v$, and the only eigenvalue of $\mathbf{P}_S - \mathbf{P}_{S'}$ is 1.

But the only eigenvalue of $\eta \eta^t / \zeta$ is $\|\eta\|_2^2 / \zeta$. By developing

$$\|\eta\|_2^2 = \|X_S(X_S^t X_S)^{-1} \varepsilon_S - X_{S'}(X_{S'}^t X_{S'})^{-1} \varepsilon_{S'}\|^2$$

and comparing with

$$\zeta = \varepsilon_S(X_S^t X_S)^{-1} \varepsilon_S - \varepsilon_{S'}^t (X_{S'}^t X_{S'})^{-1} \varepsilon_{S'},$$

we can write that the General Position Condition, Assumption 1.1, is equivalent to the following inequations:

$$\begin{aligned}\zeta &\neq 0 \\ \|\eta\|_2^2 &\neq |\zeta|.\end{aligned}$$

Therefore, the operators $\mathbf{P}_S - \mathbf{P}_{S'}$ and $\eta \eta^t / \zeta$ are different. Hence, (3.38) is not satisfied for all $y \in \mathbb{R}^n$ when the rank of $\mathbf{P}_S - \mathbf{P}_{S'}$ is 1.

As a conclusion, the dimension of E_j is less than $n + 1$. the probability that there exists $\lambda > 0$ such that uniqueness of the LASSO estimator fails, is equal to zero. \square

3.1.1. *Continuity.* Proposition 3.5 below addresses the continuity of $\widehat{\beta}_\lambda$. We start with a preliminary lemma.

Lemma 3.4. *Let Assumption 1.1 hold. Then, the function θ is concave and non-decreasing.*

Proof. Since θ is the infimum of a set of affine functions of the variable λ , it is concave. Moreover, we have

$$\theta(\lambda) = \mathcal{L}(\lambda, \widehat{\beta}_\lambda),$$

where, by Lemma 3.3, $\widehat{\beta}$ is the unique solution of (1.1). Using the filling property [8, Chapter XII], we obtain that $\partial\theta(\lambda)$ is the singleton $\{\|\widehat{\beta}_\lambda\|_1\}$. Thus, θ is differentiable and its derivative at λ is given by

$$(3.39) \quad \theta'(\lambda) = \|\widehat{\beta}_\lambda\|_1.$$

Moreover, this last expression shows that θ is nondecreasing. \square

Lemma 3.5. *Let Assumption 1.1 hold. Then, almost surely, the map*

$$\begin{cases} \mathbb{R}_+^* & \longrightarrow \mathbb{R}^p \\ \lambda & \longmapsto \widehat{\beta}_\lambda \end{cases}$$

is bounded and continuous. Moreover, its ℓ_1 -norm is non-increasing.

Proof. We naturally divide the proof into three parts:

- (i) $\|\widehat{\beta}_\lambda\|_1$ is non-increasing – The fact that $\lambda \mapsto \|\widehat{\beta}_\lambda\|_1$ is non-increasing is an immediate consequence of the concavity of θ .
- (ii) *Boundedness* – Notice that using (3.22), we obtain that

$$\|\widehat{\beta}_\lambda\|_1 \leq \max_{(S,\delta) \in \Sigma} \|(X_S^t X_S)^{-1} (X_S^t y - \lambda \delta)\|_1.$$

Thus, $\lambda \mapsto \widehat{\beta}_\lambda$ is bounded on any interval of the form $(0, M]$, with $M \in (0, +\infty)$. Moreover, since its ℓ_1 -norm is non-increasing, it is bounded on $(0, \infty)$.

- (iii) *Continuity* – Assume for contradiction that $\lambda \mapsto \widehat{\beta}_\lambda$ is not continuous at some $\lambda^\circ > 0$. Using boundedness, we can construct two sequences converging towards $\widehat{\beta}_{\lambda^\circ}^+$ and $\widehat{\beta}_{\lambda^\circ}^-$ respectively with $\widehat{\beta}_{\lambda^\circ}^+ \neq \widehat{\beta}_{\lambda^\circ}^-$. Since $\mathcal{L}(\lambda^\circ, \cdot)$ is continuous, both limits are optimal solutions of the problem

$$(3.40) \quad \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \mathcal{L}(\lambda^\circ, b),$$

hence contradicting the uniqueness proven in Lemma 3.3 above. \square

4. THE FIDELITY AND PENALTY TERMS AS FUNCTIONS OF λ

Our main goal in this section is to study the function

$$(4.41) \quad \Gamma : \begin{cases} \mathbb{R}_+ & \longrightarrow \mathbb{R}_+ \\ \lambda & \longmapsto \frac{\lambda \|\widehat{\beta}_\lambda\|_1}{\|y - X \widehat{\beta}_\lambda\|_2^2}. \end{cases}$$

This function is important in order to study the numerical aspects of the LASSO estimator. Indeed if the fidelity term is very small compared to the penalty term or vice versa, the resulting optimization problem might be very badly conditioned and the resulting estimator may turn to be useless in practice. We will prove in particular the very intuitive fact that Γ is continuous, tends to $+\infty$ when λ tends to zero and is decreasing for λ sufficiently large.

Let us first begin with the following elementary result.

Lemma 4.1. (Nontriviality of the estimator) *Let Σ be the set*

$$(4.42) \quad \Sigma = \left\{ (S, \delta); \ S \subset \{1, \dots, p\}, \ \delta \in \{-1, 1\}^{|S|}, \ |S| \leq n, \ \sigma_{\min}(X_S) > 0 \right\}.$$

The inequality

$$(4.43) \quad \inf_{(S, \delta) \in \Sigma} \|(X_S^t X_S)^{-1} (X_S^t y - \lambda \delta)\|_1 > 0$$

holds with probability one.

Proof. This is an immediate consequence of the Gaussian distribution of z . \square

Theorem 4.2. *Let Assumption 1.1 hold. Then, the function Γ defined by (4.41) almost surely satisfies*

$$(4.44) \quad \lim_{\lambda \downarrow 0} \Gamma(\lambda) = +\infty.$$

Moreover, almost surely, there exists $\tau > 0$ such that Γ is decreasing on the interval $(0, \tau]$ with $\Gamma(\tau) = 0$, while $\|y - X\hat{\beta}_\lambda\|_2$ is increasing on $(0, \tau]$.

Proof. We will use repeatedly Lemmas 3.4 and 3.5. The proof is divided into four steps.

Step 1. $\lim_{\lambda \downarrow 0} \Gamma(\lambda) = +\infty$. We divide this proof into two parts.

Step 1.a. We first show that $|\hat{T}_\lambda| = n$ for λ sufficiently small. Let $(\lambda_k)_{k \in \mathbb{N}}$ be any positive sequence converging to 0. Let β^* be any cluster point of the sequence $(\hat{\beta}_{\lambda_k})_{k \in \mathbb{N}}$ (recall that this sequence is bounded thanks to Lemma 3.5). Fix $\varepsilon > 0$ and $b \in \mathbb{R}^p$. For all $k \in \mathbb{N}$, we have

$$(4.45) \quad \mathcal{L}(\lambda_k, \hat{\beta}_{\lambda_k}) \leq \mathcal{L}(\lambda_k, b).$$

Since $\mathcal{L}(\lambda_k, \cdot)$ is continuous, we can also write for k sufficiently large:

$$\mathcal{L}(\lambda_k, \beta^*) \leq \mathcal{L}(\lambda_k, \hat{\beta}_{\lambda_k}) + \varepsilon.$$

Hence,

$$\mathcal{L}(\lambda_k, \beta^*) \leq \mathcal{L}(\lambda_k, b) + \varepsilon.$$

Letting $\lambda_k \rightarrow 0$, we obtain

$$\frac{1}{2} \|y - X\beta^*\|_2^2 \leq \frac{1}{2} \|y - Xb\|_2^2 + \varepsilon,$$

and thus,

$$(4.46) \quad \frac{1}{2} \|y - X\beta^*\|_2^2 \leq \inf_{b \in \mathbb{R}^p} \frac{1}{2} \|y - Xb\|_2^2.$$

Since $\text{range}(X) = \mathbb{R}^n$, (4.46) implies

$$\frac{1}{2}\|y - X\beta^*\|_2^2 = 0,$$

and then

$$(4.47) \quad \lim_{\lambda \downarrow 0} \|y - X\hat{\beta}_\lambda\|_2^2 = 0.$$

Notice further that $\{b \in \mathbb{R}^p, |\text{supp}(b)| < n\}$ is a finite union of subspaces of \mathbb{R}^p , each with dimension $n - 1$. Thus,

$$(4.48) \quad m := \inf_{\{b \in \mathbb{R}^p; |\text{supp}(b)| < n\}} \frac{1}{2}\|y - Xb\|_2^2 > 0,$$

with probability one. Therefore for λ sufficiently small, (4.47) implies

$$(4.49) \quad \|y - X\hat{\beta}_\lambda\|_2^2 < m,$$

from which we deduce that $|\hat{T}_\lambda| = n$.

Step 1.b. Let $\lambda_0 > 0$ be sufficiently small so that for all $\lambda \leq \lambda_0$, $|\hat{T}_\lambda| = n$. Such a λ_0 exists due to *Step 1.a.* Hence, since $X_{\hat{T}_\lambda}$ is nonsingular:

$$(4.50) \quad \mathbf{P}_{V_{T_\lambda}} = \text{Id}_n.$$

Thus, using (3.22), we obtain

$$(4.51) \quad y - X\hat{\beta}_\lambda = -\lambda X_{\hat{T}_\lambda} (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda}),$$

which implies that

$$(4.52) \quad \|y - X\hat{\beta}_\lambda\|_2^2 = \lambda^2 \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2.$$

Moreover, Lemma 4.1 combined with (3.22) gives

$$(4.53) \quad \|\hat{\beta}_\lambda\|_1 > \inf_{(S,\delta) \in \Sigma} \|(X_S^t X_S)^{-1} (X_S^t y - \lambda \delta)\|_1 > 0.$$

Hence, for $\lambda \leq \lambda_0$,

$$(4.54) \quad \Gamma(\lambda) \geq \frac{\lambda m'}{\lambda^2 \|X_{\hat{T}_\lambda} (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \gamma_{T_\lambda}\|_2^2}.$$

Using the trivial fact that

$$(4.55) \quad \sup_{(S,\delta) \in \Sigma} \|X_S (X_S^t X_S)^{-1} \delta\|_2^2 < \infty,$$

the proof is complete.

Step 2. Partitioning $(0, +\infty)$ into good intervals.

The continuity result of Lemma 3.5 implies that the interval $(0, +\infty)$ can be partitioned into subintervals of the type $I_k = (\lambda_k, \lambda_{k+1}]$, with

- (i) $\lambda_0 = 0$ and $\lambda_k \in \mathbb{R}_+^* \cup \{+\infty\}$ for $k > 0$,
- (ii) the support and sign pattern of $\hat{\beta}_\lambda$ are constant on each I_k .

Notice further that due to *Step 1.a*, $\hat{T}_\lambda \neq \emptyset$ on at least I_0 . Let \mathcal{K} be the nonempty set

$$(4.56) \quad \mathcal{K} = \left\{ k \in \mathbb{N}, \forall \lambda \in \mathring{I}_k, \hat{\beta}_\lambda \neq 0 \right\}.$$

On any interval I_k , $k \in \mathcal{K}$, Lemma 3.3 states that the expression (3.22) for $\hat{\beta}_{\hat{T}_\lambda}$ holds. Multiplying (3.22) on the left by $\text{sgn}(\hat{\beta}_{\hat{T}_\lambda})^t$, we obtain

$$(4.57) \quad \|\hat{\beta}_\lambda\|_1 = \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})^t (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} X_{\hat{T}_\lambda}^t y$$

$$(4.58) \quad -\lambda \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})^t (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda}).$$

Thus

$$\frac{d\|\hat{\beta}_\lambda\|_1}{d\lambda}(\lambda) = -\text{sgn}(\hat{\beta}_{\hat{T}_\lambda})^t (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda}),$$

on \mathbb{R}_+^* . Thus, the definition of Σ , we obtain that

$$(4.59) \quad \frac{d\|\hat{\beta}_\lambda\|_1}{d\lambda}(\lambda) \leq - \inf_{(S,\delta) \in \Sigma} \delta^t M_S^2 \delta < 0$$

on each \mathring{I}_k , $k \in \mathcal{K}$ and

$$(4.60) \quad \frac{d\|\hat{\beta}_\lambda\|_1}{d\lambda}(\lambda) = 0$$

on each \mathring{I}_k , $k \notin \mathcal{K}$, i.e. on each \mathring{I}_k such that $\|\hat{\beta}_{\hat{T}_\lambda}\|_1 = 0$ for all λ in I_k , if any such I_k exists. Since $\lambda \mapsto \|\hat{\beta}_\lambda\|_1$ is continuous on \mathbb{R}_+^* , (4.59) implies that

- (i) there exists τ in \mathbb{R}_+^* , such that $\hat{\beta}_\tau = 0$ (as an easy consequence of the Fundamental Theorem of Calculus and a contradiction).
- (ii) $\hat{\beta}_\lambda = 0$ for all $\lambda \geq \tau$.

Hence $\cup_{k \in \mathcal{K}} I_k$ is a connected bounded interval.

Step 3. $\|y - X\hat{\beta}_\lambda\|_2$ is increasing on $(0, \tau]$.

Using (4.52), we immediately see that the derivative of $\|y - X\hat{\beta}_\lambda\|_2^2$ on \mathring{I}_k is nothing but

$$(4.61) \quad \frac{d\|y - X\hat{\beta}_\lambda\|_2^2}{d\lambda} = \frac{d\lambda^2 \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2}{d\lambda}$$

$$(4.62) \quad = 2\lambda \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \text{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2.$$

Therefore,

$$(4.63) \quad \frac{d\|y - X\hat{\beta}_\lambda\|_2^2}{d\lambda} > 2\lambda n \sigma_{\min} \left((X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \right)^2 > 0,$$

which proves the desired result by using continuity of $\|y - X\hat{\beta}_\lambda\|_2$ at τ .

Step 4. Γ is decreasing on $(0, \tau)$.

Let us study the function

$$(4.64) \quad \Phi : \begin{cases} \mathbb{R}_+^* & \longrightarrow \mathbb{R}_+ \\ \lambda & \longmapsto \lambda \|\hat{\beta}_\lambda\|_1. \end{cases}$$

We immediately deduce from Step 2 and the definition of the intervals I_k , $k \in \mathcal{K}$, that Φ is differentiable on each \mathring{I}_k , $k \in \mathcal{K}$, and using (3.22), its derivative on \mathring{I}_k reads

$$(4.65) \quad \begin{aligned} \frac{d\Phi}{d\lambda}(\lambda) &= \|\hat{\beta}_{\hat{T}_\lambda}\|_1 - \lambda \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})^t (X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda}) \\ &= \|\hat{\beta}_{\hat{T}_\lambda}\|_1 - \lambda \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1/2} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2. \end{aligned}$$

Now, since $X_{\hat{T}_\lambda}$ is non singular,

$$(4.66) \quad \|y - X\hat{\beta}_\lambda\|_2^2 = \lambda^2 \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2$$

$$(4.67) \quad > \lambda^2 n \sigma_{\min}((X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1})^2 > 0$$

for $\lambda > 0$. Therefore $\Gamma(\lambda) < +\infty$ on \mathbb{R}_+^* , Γ is continuous on I_k and differentiable on \mathring{I}_k . Moreover, using (4.52), we have

$$\begin{aligned} \frac{d\Gamma}{d\lambda}(\lambda) &= \frac{\frac{d\Phi}{d\lambda}(\lambda) \|y - X\hat{\beta}_\lambda\|_2^2 - \Phi(\lambda) \frac{d\|y - X\hat{\beta}_\lambda\|_2^2}{d\lambda}(\lambda)}{\|y - X\hat{\beta}_\lambda\|_2^4} \\ &= \frac{\frac{d\Phi}{d\lambda}(\lambda) - 2 \frac{\Phi(\lambda)}{\lambda}}{\|y - X\hat{\beta}_\lambda\|_2^2}. \end{aligned}$$

Hence, using (4.65) and (4.52),

$$\begin{aligned} \frac{d\Gamma}{d\lambda}(\lambda) &= \frac{-\|\hat{\beta}_{\hat{T}_\lambda}\|_1 - \lambda \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1/2} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2}{\|y - X\hat{\beta}_\lambda\|_2^2} \\ &\leq \frac{-\lambda \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1/2} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2}{\lambda^2 \|(X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1} \operatorname{sgn}(\hat{\beta}_{\hat{T}_\lambda})\|_2^2} \\ &\leq -\frac{1}{\lambda} \left(\frac{\sigma_{\min}((X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1/2})}{\sigma_{\max}((X_{\hat{T}_\lambda}^t X_{\hat{T}_\lambda})^{-1})} \right)^2, \end{aligned}$$

on each \mathring{I}_k . We can thus conclude, by the non-singularity of $X_{\hat{T}_\lambda}$ that Γ is decreasing on $(0, \tau)$, as announced. \square

REFERENCES

1. Bickel, P. J., Ritov, Y., Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (2009), no. 4, 1705–1732.
2. Becker, S., Bobin, J. and Candès, E. J., NESTA: a fast and accurate first-order method for sparse recovery. In press SIAM J. on Imaging Science.
3. Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Sparsity oracle inequalities for the Lasso. Electron. J. Stat., 1 :169–194.

4. Bunea, F., Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization, the Electronic Journal of Statistics, (2008) Vol. 2, 1153-1194
5. Candès, E. J. Modern statistical estimation via oracle inequalities. Acta Numer. 15 (2006), 257–325.
6. Candès, E. J. and Plan, Yaniv. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 285–2177.
7. Dossal, C., A necessary and sufficient condition for exact recovery by ℓ_1 minimization. <http://hal.archives-ouvertes.fr/docs/00/16/47/38/PDF/DossalMinimisation11.pdf>
8. Hiriart-Urruty, J.-B. and Lemaréchal, C. Convex Analysis and Minimization Algorithms II. Advanced theory and bundle methods. Grundlehren der Mathematischen Wissenschaften 306. Springer Verlag.
9. E. del Barrio, P. Deheuvels and S.A. van de Geer (2007). Lectures on Empirical Processes. EMS Series of Lectures in Mathematics, European Mathematical Society Publishing House (2007)
10. Schrijver, A., Theory of linear and integer programming. Wiley-Interscience Series in Discrete Mathematics. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, 1986. xii+471 pp.
11. Tibshirani, R. Regression shrinkage and selection via the LASSO, J.R.S.S. Ser. B, 58, no. 1 (1996), 267–288.
12. van de Geer, S., High-dimensional generalized linear models and the Lasso. The Annals of Statistics 36, 614-645.

LABORATOIRE DE MATHÉMATIQUES, UMR 6623, UNIVERSITÉ DE FRANCHE-COMTÉ,
 16 ROUTE DE GRAY,, 25030 BESANCON, FRANCE
E-mail address: `stephane.chretien@univ-fcomte.fr`

LATP, UMR 6632, UNIVERSITÉ DE PROVENCE, TECHNOPÔLE CHÂTEAU-GOMBERT,
 39 RUE JOLIOT CURIE, 13453 MARSEILLE CEDEX 13, FRANCE
E-mail address: `darses@cmi.univ-mrs.fr`